Do Large Language Models *Think* Like the Brain? Sentence-Level Evidence from fMRI and Hierarchical Embeddings

Yu Lei[1*‡], Xingyang Ge[2,8*], Yi Zhang[3,4], Yiming Yang [6,7,8†], Bolei Ma[4,5†]

[1]Beijing University of Posts and Telecommunications [2]Shandong University
[3]FAU Erlangen-Nuremberg [4]LMU Munich [5]Munich Center for Machine Learning
[6]Linguistic Science Laboratory, Jiangsu Normal University
[7]Collaborative Innovation Center for Language Ability, Jiangsu Normal University
[8]School of Linguistic Sciences and Arts, Jiangsu Normal University

*These authors contributed equally, †Corresponding authors, ‡Project lead

leiyu@bupt.edu.cn, 202320175@mail.sdu.edu.cn, yi.zhang@fau.de, yangym@jsnu.edu.cn, bolei.ma@lmu.de

Abstract

Understanding whether large language models (LLMs) and the human brain converge on similar computational principles remains a fundamental and important question in cognitive neuroscience and AI. Do the brain-like patterns observed in LLMs emerge simply from scaling, or do they reflect deeper alignment with the architecture of human language processing? This study focuses on the sentence-level neural mechanisms of language models, systematically investigating how hierarchical representations in LLMs align with the dynamic neural responses during human sentence comprehension. By comparing hierarchical embeddings from 14 publicly available LLMs with fMRI data collected from participants, who were exposed to a naturalistic narrative story, we constructed sentence-level neural prediction models to precisely identify the model layers most significantly correlated with brain region activations. Results show that improvements in model performance drive the evolution of representational architectures toward brain-like hierarchies, particularly achieving stronger functional and anatomical correspondence at higher semantic abstraction levels.

1 Introduction

The intersection of artificial intelligence and neuroscience has emerged as a cutting-edge research frontier, particularly in understanding the parallels between large language models (LLMs) and human neural language processing (Toneva and Wehbe, 2019; Abnar et al., 2019; Schrimpf et al., 2021). Prior studies (Anderson et al., 2021; Caucheteux et al., 2021) have established foundational evidence showing intriguing correlations between LLM-learned representations and neural responses during language processing, particularly in feature extraction and representational similarity (Caucheteux and King, 2022; Hosseini et al., 2024). These findings (Sun et al., 2021) suggest that both systems may utilize comparable linguistic features, as evidenced by the linear mappability of LLM representations to neural activity patterns. However, these observations lack mechanistic explanations for what critical properties enable LLMs to achieve brain-like processing capabilities.

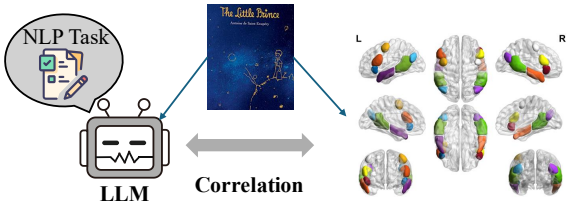

Figure 1: A brief presentation of the experimental design. We expose both LLMs and real humans to a narrative story ("The Little Prince") and aim to compare the correlation between LLMs' and human brains' language processing.

Recent investigations (Goldstein et al., 2022; Caucheteux et al., 2023) have explored multimodal similarities between LLMs and neural processes. While some studies (Antonello et al., 2023; Antonello and Huth, 2024) demonstrate stronger alignment between autoregressive LLMs (Ethayarajh, 2019; Tenney et al., 2019) and the predictive coding hypothesis of human language processing, others focus on metrics like language modeling performance, model scale, and representational generalization. Collectively, these works indicates that modeling quality critically determines brain-like representational capacities (Hickok and Poeppel, 2007; Hasson et al., 2008; Lerner et al., 2011; Ding et al., 2017). A fundamental question remains unresolved: *Does this similarity merely stem from increased model scale, or does it reflect deeper*

convergence in computational principles with the human speech processing pathway? Resolving this dichotomy is crucial for advancing next-generation model architectures.

Current studies examining the relationship between LLMs and brain decoding typically use publicly available datasets to evaluate model performance in open scenarios. However, such datasets may fail to accurately reflect the models' comprehension abilities in specific tasks. To address this limitation, we selected 14 publicly available pretrained LLMs and designed a sentence understanding task to assess their contextual understanding. Additionally, we utilized fMRI data from participants who listened to naturalistic text to construct decoding models, using the *corr* metric to compare the relationship between LLMs and brain-activity patterns in relevant regions. A simple presentation of our experimental design is illustrated in Figure 1

We summarize our main findings as follows:

1) **Performance on the sentence understanding task:** Instruction-tuned models consistently outperformed base models (§4.1).

2) **Correlation with brain activations:** Across all LLMs, middle layers achieved the best performance in predicting brain activation, and we found a significant positive association between model performance and brain correlations. Instruction-tuned models exhibited higher correlation values than their corresponding base models, with this difference demonstrating statistically significant correlation. (§4.2).

3) **Hemispheric asymmetry and functional specificity:** Specific brain regions exhibited left-hemispheric lateralization, likely enhancing processing efficiency through specialized neural mechanisms. Additionally, hemispheric asymmetry in other brain regions revealed function-specific contributions to the models' performance, indicating differentiated functionality across regions (§4.3).

2 Related Work

Neuroscientific Foundations of Language Comprehension. Before the advent of current large-scale LLMs, cognitive neuroscience had established that human language comprehension relies on complex hierarchical processing (Friederici, 2011). Research centered on understanding how the brain integrates perceptual units (e.g., phonemes and graphemes) into meaningful structures at the lexical, sentential, and discourse levels (Price, 2012). Early models, such as the Wernicke-Lichtheim-Geschwind model (Geschwind, 1967), identified localized brain regions for language but struggled to explain sentence-level and discourse-level processing, especially information integration across sentences for coherence (Hickok and Poeppel, 2004). Neuroimaging techniques like fMRI and ERP (Kutas and Hillyard, 1984; Osterhout and Holcomb, 1992) led to a shift from localized to distributed network models, such as the Memory-Unification-Control (MUC) framework (Hagoort, 2016). However, early fMRI studies were limited by isolated stimulus presentation and signal averaging, which made capturing long-range linguistic integration difficult (Humphries et al., 2007). New naturalistic paradigms like narrative listening (Yarkoni et al., 2008; Brennan, 2016) advanced investigation into discourse comprehension, but approaches like representational similarity analysis (RSA) (Kriegeskorte et al., 2008) often underestimate core discourse features like coherence and context-dependent meaning (Zacks et al., 2017; Xu et al., 2024; Messi and Pylkkanen, 2025), emphasizing the need for multi-sentence integration analyses.

Leveraging LLMs for Brain-Language Mapping. Recent studies (Luo et al., 2022; Yu et al., 2024) use the strong semantic capabilities of pretrained language models to examine brain-language mappings and decode neural processes. For instance, (Ren et al., 2025; Yu et al., 2024) applied Dynamic Similarity Analysis (DSA) to compare text embeddings with fMRI signals, constructing Representational Dissimilarity Matrices (RDMs) using measures like Pearson correlation. Other works (Mischler et al., 2024; Bonnasse-Gahot and Pallier, 2024) aligned layer-wise activations of language models with averaged fMRI activity maps using ridge regression. Moreover, (Tuckute et al., 2024) trained encoding models on fMRI data from participants exposed to diverse sentences, optimizing GPT-2 XL embeddings (specifically the 22nd layer) for neural alignment. These studies showcase the potential of LLMs to reveal insights into the neural mechanisms of language comprehension, offering new tools and methodologies for cognitive neuroscience.

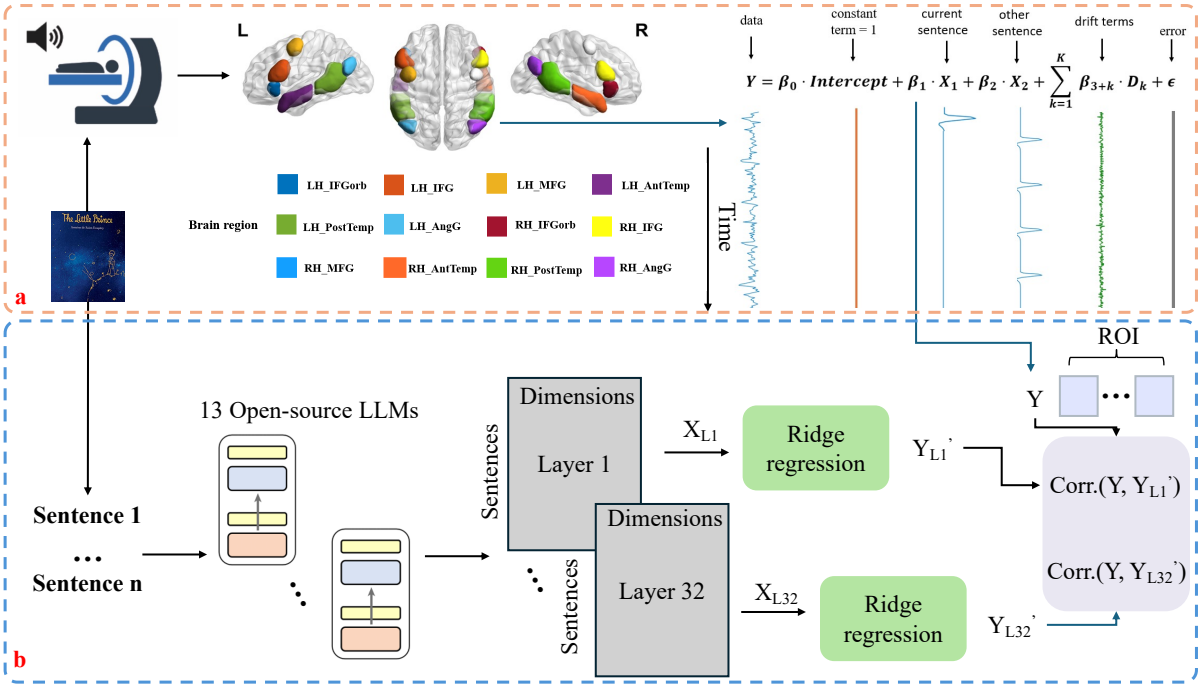

Figure 2: This figure illustrates the multi-stage pipeline used to analyze the alignment between LLM representations and neural responses during naturalistic language comprehension. The methodology includes auditory stimulus presentation, hierarchical embedding extraction from LLMs, voxel-wise regression modeling, and region-of-interest (ROI)-based brain-model alignment analysis. Panel (a) outlines the neuroimaging data acquisition and preprocessing steps (§3.1 to §3.4), while panel (b) describes the brain-LLM alignment analysis (§3.5).

3 Methodology

To systematically examine the alignment between LLM representations and neural responses during naturalistic language comprehension, we implement a multi-stage pipeline combining neuroimaging data analysis, model feature extraction, and encoding model evaluation. The overall methodology is summarized in Figure 2. We present in the following the detailed experimental methods, which outline the integration of auditory stimulus presentation, LLM-based hierarchical embedding extraction, voxel-wise regression modeling, and region-of-interest (ROI)-based brain-model alignment analysis. The detailed brain data acquisition and processing and fMRI data extraction pipeline is documented from §3.1 to §3.4, as shown in Figure 2(a). Then, the brain-LLM alignment analysis is detailed in §3.5, as shown in Figure 2(b).

3.1 Data and Stimuli

We used the existing data from (Li et al., 2022) which provides the "The Little Prince" multilingual naturalistic fMRI corpus. In their dataset, 35 healthy, right-handed, young native Mandarin speakers (anonymous) participated in the experiment (15 females; mean age = 19.3 years, range = 18–25, One participant was excluded due to partial missing fMRI functional data). All participants had no history of psychiatric disorders, neurological diseases, or other conditions that might affect cognitive function. They resided in China and had studied English since middle school.

The stimuli used are parallel corpora of the diary "The Little Prince" in both English and Chinese. The texts are segmented into 1,577 sentences in Chinese, with corresponding English sentences aligned for each one. These parallel sentence pairs are also used for the analyses in Section 4.1.

3.2 fMRI Acquisition and Preprocessing

After familiarizing themselves with the MRI environment, participants lay supine in the scanner. Experimental scripts were presented via PsychoPy 2 (Peirce et al., 2019), and auditory stimuli were delivered through MRI-compatible high-fidelity headphones. The Chinese audiobook material totaled approximately 99 minutes and was divided into 9 segments (each 10 minutes). After passively listening to each segment, participants completed 4 comprehension questions per segment (36 questions total). Questions were viewed via a mirror attached to the head coil, and responses were recorded using

a button box. The entire procedure lasted approximately 2.5 hours.

Data were acquired using a 3T GE Discovery MR750 scanner with a 32-channel head coil. Structural images were obtained via a T1-weighted MP-RAGE sequence, and functional images were acquired using a multi-echo planar imaging (ME-EPI) sequence (TR = 2000 ms; TEs = 12.8, 27.5, 43 ms; voxel size = 3.75 × 3.75 × 3.8 mm). Preprocessing was performed using AFNI 16 (Cox, 1996), including removal of the first 4 time points, multi-echo independent component analysis (ME-ICA) for denoising, and spatial normalization to the MNI standard space (resampled to 2 mm³ voxels).

3.3 General Linear Model (GLM)

The General Linear Model (GLM) serves as the gold-standard analytical framework in fMRI neuroscience research, utilizing statistical modeling to precisely isolate task-evoked blood-oxygen-level-dependent (BOLD) signals from physiological and system noise. The core model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{1}$$

where $\mathbf{Y}$ represents the BOLD time series, $\mathbf{X}$ is the design matrix encoding experimental conditions and nuisances, $\boldsymbol{\beta}$ contains regression coefficients, and $\boldsymbol{\epsilon}$ is residual noise.

The design matrix $\mathbf{X}$ is constructed by convolving event onsets with a canonical hemodynamic response function (HRF). Each condition's regressor $\mathbf{x}_j$ includes the trial onsets and nuisance confounds (e.g., motion, drift):

$$\mathbf{x}_j(t) = \sum_{k=1}^{K_j} \text{HRF}(t - t_{\text{onset}}^{(j,k)}) + \sum_{m=1}^{M} \gamma_m \mathbf{n}_m(t) \tag{2}$$

We adopt the Least-Squares Separate (LS-S) approach (Mumford et al., 2014) to isolate BOLD responses for each sentence-level trial by modeling each target sentence independently. This method reduces collinearity and improves sensitivity to transient, event-related activity compared to block-based or condition-averaged models. LS-S enables precise sentence-level neural activation patterns under naturalistic stimuli.

Voxel-wise parameter estimates are obtained as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \tag{3}$$

t-statistics are computed for testing contrasts $\mathbf{c}$:

$$t_v = \frac{\mathbf{c}^\top\hat{\boldsymbol{\beta}}_v}{\sqrt{\hat{\sigma}_v^2 \cdot \mathbf{c}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{c}}}, \quad \hat{\sigma}_v^2 = \frac{\|\mathbf{Y}_v - \mathbf{X}\hat{\boldsymbol{\beta}}_v\|^2}{T - \text{rank}(\mathbf{X})} \tag{4}$$

This enables inference on condition-specific activations while accounting for noise and drift, improving sensitivity and interpretability of neural responses to linguistic stimuli.

3.4 Region of Interest (ROI) Extraction

We then need to extract the ROIs. We employ the language network proposed by (Fedorenko et al., 2010) as its core analytical framework to investigate neurobiological distinctions between human brains and LLMs in natural language comprehension. Widely validated as the neural substrate for representing linguistic knowledge and supporting core language computations - including lexical access, syntactic parsing, and semantic composition (Tuckute et al., 2024). Following the implementation of the GLM, ROIs will be utilized to systematically extract condition-specific beta values through predefined anatomical atlas masks. These beta coefficients quantify the response magnitude of BOLD signals under specific experimental conditions (individual sentence stimuli). The detailed functions of each ROI and their significance to human language processing are documented in Appendix §A (Table2).

3.5 Cross-Validated Ridge Regression for Layer-wise Encoding Analysis

We then introduce the LLM processing of "The Little Prince", as shown in Figure 2(b). The text is segmented into sentences and then fed into the LLMs. The details of the LLMs are documented in Appendix E.

In this framework, we employ ridge regression to quantify how semantic representations from different layers of a neural network relate to brain activity patterns captured via fMRI. By integrating deep learning and neuroscience, we systematically evaluate how well neural embeddings predict fMRI signals for a specific region of interest (ROI). Given fMRI response vectors $\mathbf{y} \in \mathbb{R}^N$ and neural embeddings $\mathbf{X} \in \mathbb{R}^{N \times L \times D}$ (where T is the number of time points, L is the number of layers, and D denotes embedding dimensionality), the predictive performance for each layer is computed as the average Pearson correlation (ρ) across K cross-validation folds:

$$\rho_l = \frac{1}{K}\sum_{k=1}^{K} \text{corr}\left(\mathbf{y}_{\text{test}}^{(k)}, \mathbf{X}_{\text{test}}^{(l,k)}\hat{\boldsymbol{\beta}}^{(l,k)}\right) \tag{5}$$

The model uses ridge regression to estimate these predictions, balancing data fidelity and regu-

larization to avoid overfitting. Regression weights ($\boldsymbol{\beta}$) are optimized with:

$$\hat{\boldsymbol{\beta}}^{(l,k)} = (\mathbf{X}_{\text{train}}^{(l,k)\top}\mathbf{X}_{\text{train}}^{(l,k)} + \alpha\mathbf{I})^{-1}\mathbf{X}_{\text{train}}^{(l,k)\top}\mathbf{y}_{\text{train}} \quad (6)$$

Here, α controls the regularization strength, selected via grid search in a nested cross-validation framework. All data are z-scored to ensure compatibility between fMRI and neural embeddings.

Feature Normalization and Parallelization. Both $\mathbf{y}$ and $\mathbf{X}$ are standardized to zero mean and unit variance before regression to improve numerical stability. The analysis pipeline parallelizes computations across subjects, ROIs, and layers, training $\mathcal{O}(SRL)$ models, where S is the number of subjects, R is the number of ROIs, and L is the number of layers. This ensures efficient scalability in the layer-wise evaluation of embeddings.

This approach provides layer-wise correlation metrics (ρ_l), which can offer insights into how neural network representations align with ROI-specific brain activity. We refer to the details of the ridge regression process in Appendix §C

4 Experiments and Results

4.1 Performance of Large Language Models

To evaluate the contextual understanding capabilities of LLMs under cross-lingual settings, we first design a multiple-choice test based on semantic alignment between Chinese source sentences and their English translations. Drawing from recent work on evaluating the robustness of LLMs in multiple choice setups (Zheng et al., 2024; Wang et al., 2024), we adopt several perturbations to test the LLMs' performance. For each original Chinese sentence from the stimuli, we generate five English options as shown in Table 1.

Option	Example Sentence
A	it showed a boa constrictor swallowing a wild animal
B	**animal wild** showed a boa **it swallowing** a constrictor
C	it showed a boa constrictor swallowing a **wild beast**
D	**A wild animal was shown being swallowed** by a boa constrictor
E	it showed a boa constrictor swallowing a **briefcase in a busy city**

Table 1: Examples of English translation options demonstrating five types of semantic variation: (A) Correct Translation, (B) Word Order Scrambled, (C) Part-of-speech Substitution, (D) Sentence Structure Transformation, (E) Information Insertion/Deletion. Details of the variation description are presented in Appendix D.

Each of these English sentence options is then input into the LLM to obtain their corresponding embeddings. This experimental design allows us to systematically analyze the LLM's ability to distinguish the correct translation from various types of distractors and to assess its cross-lingual semantic alignment performance.

To quantitatively evaluate the cross-lingual semantic understanding capability of large language models, we propose the **Cross-lingual Semantic Alignment Accuracy (CSAA)** metric. This metric is defined as the proportion of cases in which the model correctly identifies the true translation (option A) as the most semantically similar candidate to the original Chinese sentence, as measured by cosine similarity in the embedding space.

We define an indicator variable δ_i for each sample i as follows:

$$\delta_i = \begin{cases} 1, & \text{if } \operatorname{argmax}_x \left(\cos\left(\mathbf{v}_{c_i}, \mathbf{v}_{e_{i,x}}\right)\right) = A \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

The CSAA is then defined as:

$$\text{CSAA} = \frac{1}{N}\sum_{i=1}^{N}\delta_i \quad (8)$$

where N is the total number of Chinese sentences, $\mathbf{v}_{c_i}$ denotes the embedding of the i-th Chinese sentence, and $\mathbf{v}_{e_{i,x}}$ denotes the embedding of the x-th English candidate for the i-th sentence. A higher CSAA indicates stronger cross-lingual semantic alignment ability of the model.

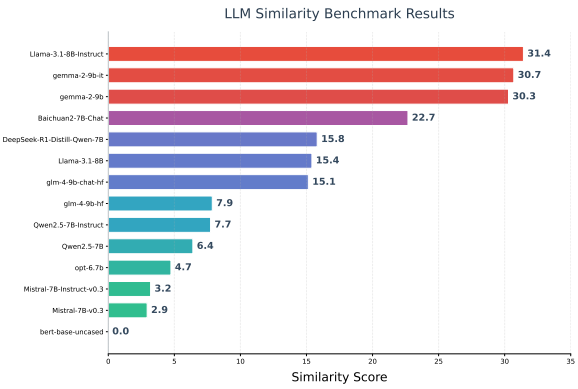

Figure 3: Performance comparison of 14 LLMs

The results in Figure 3 show that the Llama-3.1-Instruct version leads with 31.4 points, followed by two Gemma variants (30.7 and 30.3). A sharp performance drop occurs after the top three models, with mid-tier scores (22.7-15.1) dominated by Baichuan2 and DeepSeek variants, while glm-4.9b (7.9) and Qwen2.5-7B (6.4) anchor the lower end. The red-to-green gradient visually reinforces score

Figure 4: All correlation between model predictions and brain activity across layers. Shaded areas represent the 95% confidence intervals.

disparities, highlighting a >28-point gap between best and worst performers. Consistently, we notice that the instruction-tuned models always perform better than the base models in each model family.

4.2 Model Performance and Activation Correlation

To comprehensively assess the correspondence between computational models and brain activity, we compared model performance across multiple brain regions and layers.

The comparative evaluation of various computational models across all brain regions is summarized in Figure 4, which displays the layer-wise correlation coefficients together with their 95% confidence intervals. Each curve represents the average performance trajectory of a given model, with shaded regions indicating confidence bounds. Notably, we found that in most cases, models achieve peak predictive performance at their intermediate layers, not the final layer, consistent with previous findings (Mischler et al., 2024).

To further compare models, we selected the optimal layer exhibiting peak performance as the model's output representation and examined the average correlation across 12 ROIs, as shown in Figure 5. More detailed results for individual ROIs can be found in Appendix Figure 9. Our results show that LLMs consistently yield higher correlation metrics than the base BERT model across

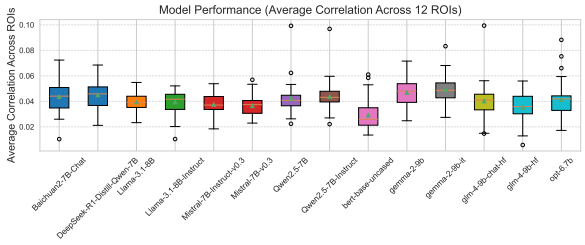

Figure 5: Average correlation of LLMs across 12 ROIs.

brain regions.

From the 14 LLMs, we selected 5 instruction-tuned models (which have their corresponding base versions) along with their 5 corresponding base versions for comparative analysis. The selection was motivated by the intent to directly compare the impact of instruction tuning. As shown in Figure 6, instruction-tuned models exhibit performance improvements in both Correlation and Performance metrics when compared with the base models. Specifically, the p-value for the Correlation Change indicates a trend toward significance, while the Performance Change (permutation test, $p = 0.03125$) demonstrates that the observed performance gains are not only substantial in magnitude but also consistent enough to reach statistical significance. These results underscore the potential of instruction-tuning in enhancing model behavior.

We present he relationship between model performance and layer activation correlations in Fig-

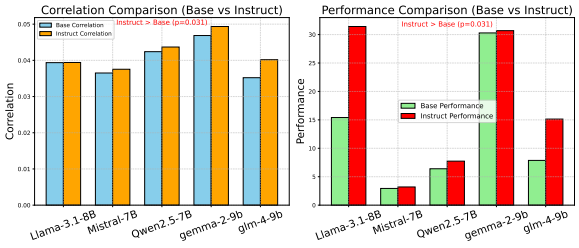

Figure 6: Comparison of instruction-tuned versus base models across performance and activation correlations

ure 7, where higher performance scores show weak positive correlation with activation values across 6.7B to 9B parameter models (Pearson correlation: $r = 0.601*$, $p = 0.030$). Notable examples include Llama-3-1-8B-Instruct (6.7B) and gemma-2.9b (9B), with activation patterns demonstrating parameter-scale dependent clustering (pink-to-purple gradient). The scatter plot's white grid background and annotated confidence interval (light blue shading) enhance comparative analysis of model architectures.

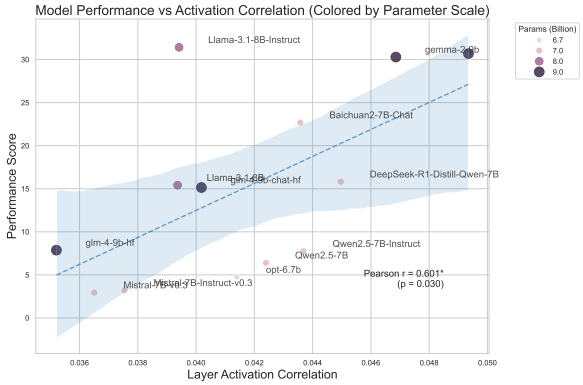

Figure 7: Correlation between model performance and activation patterns across LLMs

4.3 Left-right Hemispheric Asymmetry

In this section, we investigate the lateralization patterns of neural activity associated with LLM (6.7B–9B parameters) processing across key language-related brain regions, focusing on the asymmetry between left and right hemispheres and its relationship to model performance. In Figure 8(a), we visualize the localization of ROIs in both hemispheres.

We observed left-right hemispheric asymmetry in neural activity correlations between specific brain ROIs and LLMs. In Figure 8(b), the inferior frontal gyrus (IFG) and posterior temporal (PostTemp) regions showed stronger left-hemispheric dominance in LLM-related neural correlations, consistent with their roles in core language functions like production and comprehension (Hu et al., 2023). Conversely, the middle frontal gyrus (MFG) and anterior temporal (AntTemp) regions exhibited stronger right-hemispheric involvement, possibly reflecting specialization in tasks such as metaphor processing, contextual integration, and cross-modal semantics. Right MFG correlations may relate to LLMs' demands for cognitive control in complex narratives (Japee et al., 2015), while right AntTemp activity likely supports multimodal semantic representations (One-Sample t-test: IFG: $p = 0.025$; PostTemp: $p = 0.007$; MFG: $p = 0.005$; AntTemp: $p = 0.001$). These findings align with the classical language lateralization hypothesis of left-hemisphere dominance for syntax/semantics and emerging evidence of right-hemisphere contributions to cognitive control (Vigneau et al., 2006; Menon and D'Esposito, 2022).

To examine the relationship between hemispheric asymmetry and model performance, Figure 8(c) shows correlations between LH-RH differences and performance metrics. Among the six ROIs analyzed, IFG and MFG lateralization were potentially linked to performance, showing positive relationships (Pearson correlation: IFG: $r = 0.54$, $p = 0.055$; MFG: $r = 0.50$, $p = 0.084$). No significant correlations were found for other regions(Pearson correlation: AntTemp: $r = 0.02$, $p = 0.941$; PostTemp: $r = 0.09$, $p = 0.770$; AngG: $r = 0.09$, $p = 0.687$; IFGorb: $r = -0.06$, $p = 0.845$). The regulatory role of the prefrontal cortex over distributed brain regions may underlie this phenomenon (Badre and Nee, 2018), where its lateralized functional specialization could enhance top-down coordination of cognitive resources, mirroring the efficiency optimization observed in computational models.

5 Discussion

Understanding how LLMs align with human brain activity during sentence processing provides critical insights into both artificial and biological language systems (Mahowald et al., 2024; Zhou et al., 2024).

From the perspective of brain decoding in sentence tasks, our model also exhibits higher activation correlations in the middle layers compared to the final layer, which is consistent with the findings of (Mischler et al., 2024). Importantly, our

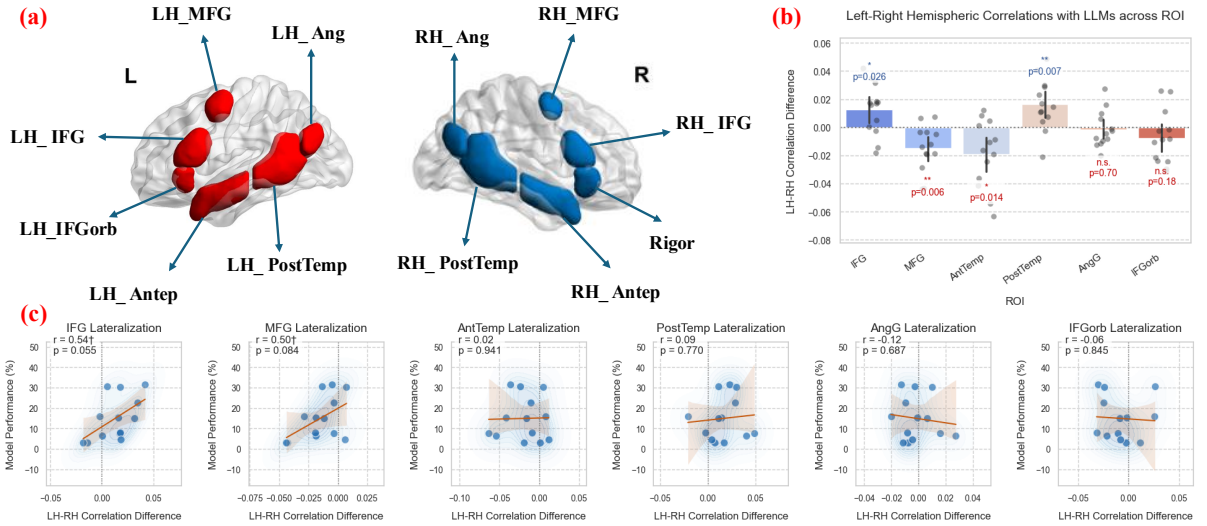

Figure 8: (a) illustrates the localization of ROIs in both hemispheres (left hemisphere: red; right hemisphere: blue); (b) displays the left-minus-right (LH-RH) correlation differences in ROI-LLM associations; (c) examines the relationship between LH-RH asymmetry and model performance.

exploration using fMRI data complements their approach based on intracranial electrode recordings and MEG (Zhou et al., 2024). At the level of sentence processing, Yu et al. (2024) investigated the process from the perspective of BERT, whereas we not only consider BERT but also compare it with larger LLMs (6.7B–9B parameters). We further find that instruction-tuned models outperform base models in both correlation and sentence comprehension ability, with marginally significant improvements in correlation. This observation is in line with (Ren et al., 2025), which also reported that instruction tuning can significantly enhance LLM-brain similarity. In addition, while previous studies such as (Bonnasse-Gahot and Pallier, 2024) primarily focus on the relationship between model size and brain correlation, our work shifts the emphasis to the model's comprehension ability. Specifically, within the parameter range of 6.7B to 9B, we reveal not only a significant positive association between comprehension ability and neural similarity, but also identify a scaling law: as the comprehension ability of language models increases, their neural alignment with brain activity systematically and predictably improves.

Left-right hemispheric asymmetry is a fundamental organizational principle of the human brain, shaping the specialization of language and cognitive functions. The left-hemispheric dominance in IFG and PostTemp aligns with their established roles in core language functions (Hu et al., 2023), supporting syntactic encoding and semantic integration. In contrast, right-hemispheric engagement in MFG and AntTemp may reflect specialized processes for metaphor comprehension, contextual integration (Japee et al., 2015), and multimodal semantic representations. The positive trends between leftward IFG/MFG asymmetry and model performance suggest that enhanced neural specialization in these regions may reduce inter-hemispheric interference and improve processing efficiency (Cai et al., 2013; Friedman and Robbins, 2022). Stronger left IFG activation has been previously linked to superior language capabilities (Gotts et al., 2013; Arredondo et al., 2019). Non-significant correlations in other regions may indicate bilateral coordination or domain-general processes, such as the AngG's role in the default mode network and non-lateralized semantic integration (Kuhnke et al., 2023).

These findings highlight that LLM processing engages both conserved left-hemispheric language systems and right-hemispheric networks supporting higher-order cognitive demands. The functional specificity of hemispheric asymmetries underscores the need to disentangle distinct neural contributions to LLM performance, rather than attributing it to a singular mechanism.

6 Conclusion

This study bridges LLMs and human brain activity by introducing a custom sentence understanding task paired with naturalistic fMRI data. Our

evaluation of 14 open LLMs reveals that instruction tuning substantially enhances both task performance and alignment with neural responses, demonstrating statistically significant correlations. Additionally, we observe that the performance of the intermediate layers surpasses that of the final layer within the LLMs. Moreover, we identify clear patterns of left-hemispheric dominance in key language regions, suggesting efficient, specialized processing, alongside function-specific asymmetries that reflect differentiated neural roles in supporting language comprehension. Together, these findings demonstrate that model architecture and training objectives shape the degree of brain alignment, and highlight the value of integrating cognitively grounded tasks with neuroimaging to better understand the neural basis of language and guide future model development.

Limitations

This study was limited to neurotypical participants, without inclusion of individuals with language impairments or neurodevelopmental diversity. As a result, the clinical relevance of the findings is constrained, as atypical language processing patterns in populations such as individuals with aphasia or dyslexia were not captured. Additionally, the participant sample was culturally homogeneous, consisting primarily of native Chinese speakers, which limits the generalizability of the results to cross-linguistic or bilingual populations. These constraints may affect the applicability of our conclusions, particularly in clinical or cross-cultural contexts, and highlight the need for more inclusive sampling in future research.

Ethics Statements

The fMRI data we used in the current study are from a publicly available resource (Li et al., 2022). The current study was reviewed by the IRB of Jiangsu Normal University. The study itself does not raise any ethical concerns.

Acknowledgments

The authors acknowledge the use of ChatGPT exclusively for grammatical checks in the final manuscript.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Andrew James Anderson, Douwe Kiela, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Edmund C. Lalor. 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, 41(18):4100–4119.

Richard Antonello and Alexander Huth. 2024. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79.

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2023. Scaling laws for language encoding models in fmri. In *Advances in Neural Information Processing Systems*, volume 36, pages 21895–21907. Curran Associates, Inc.

Maria M Arredondo, Xiao-Su Hu, Erica Seifert, Teresa Satterfield, and Ioulia Kovelman. 2019. Bilingual exposure enhances left ifg specialization for language in children. *Bilingualism: Language and Cognition*, 22(4):783–801.

David Badre and Derek Evan Nee. 2018. Frontal cortex and the hierarchical control of behavior. *Trends in cognitive sciences*, 22(2):170–188.

Laurent Bonnasse-Gahot and Christophe Pallier. 2024. fmri predictors based on language models of increasing complexity recover brain left lateralization. *arXiv preprint arXiv:2405.17992*.

Jonathan Brennan. 2016. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Qing Cai, Lise Van der Haegen, and Marc Brysbaert. 2013. Complementary hemispheric specialization for language production and visuospatial attention. *Proceedings of the National Academy of Sciences*, 110(4):E322–E330.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1336–1348. PMLR.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134. © 2022. The Author(s). Research Support, Non-U.S. Gov't.

Nai Ding, Lucia Melloni, Aotian Yang, Yu Wang, Wen Zhang, and David Poeppel. 2017. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (eeg). *Frontiers in human neuroscience*, 11:481.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New method for fmri investigations of language: Defining rois functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194. PMID: 20410363.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. 2024. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312.

Evelina Fedorenko, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262.

Angela D Friederici. 2011. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392.

Naomi P Friedman and Trevor W Robbins. 2022. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, 47(1):72–89.

Norman Geschwind. 1967. Wernicke's contribution to the study of aphasia. *Cortex*, 3(4):449–463.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Stephen J Gotts, Hang Joon Jo, Gregory L Wallace, Ziad S Saad, Robert W Cox, and Alex Martin. 2013. Two distinct forms of functional lateralization in the human brain. *Proceedings of the National Academy of Sciences*, 110(36):E3435–E3444.

Peter Hagoort. 2016. Muc (memory, unification, control): A model on the neurobiology of language beyond single word processing. In *Neurobiology of language*, pages 339–347. Elsevier.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. 2008. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.

Gregory Hickok and David Poeppel. 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63.

Jennifer Hu, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zekelman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, and Evelina Fedorenko. 2023. Precision fmri reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, 33(8):4384–4404.

Colin Humphries, Jeffrey R Binder, David A Medler, and Einat Liebenthal. 2007. Time course of semantic processes during sentence comprehension: an fmri study. *Neuroimage*, 36(3):924–932.

Shruti Japee, Kelsey Holiday, Maureen D Satyshur, Ikuko Mukai, and Leslie G Ungerleider. 2015. A role of right middle frontal gyrus in reorienting of attention: a case study. *Frontiers in systems neuroscience*, 9:23.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Philipp Kuhnke, Curtiss A Chapman, Vincent KM Cheung, Sabrina Turker, Astrid Graessner, Sandra Martin, Kathleen A Williams, and Gesa Hartwigsen. 2023. The role of the angular gyrus in semantic cognition: a synthesis of five functional neuroimaging studies. *Brain Structure and Function*, 228(1):273–291.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, 9(1):530.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvincé Pongos, Idan A. Blank, Melissa Kline Struhl, Anna Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castañón, and Evelina Fedorenko. 2022. Probabilistic atlas for the language network based on precision fmri data from >800 individuals. *Scientific Data*, 9(1):529.

Yifei Luo, Minghui Xu, and Deyi Xiong. 2022. CogTaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 904–920, Dublin, Ireland. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.

Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019.

Vinod Menon and Mark D'Esposito. 2022. The role of pfc networks in cognitive control and executive function. *Neuropsychopharmacology*, 47(1):90–103.

Aline-Priscillia Messi and Liina Pylkkanen. 2025. Tracking neural correlates of contextualized meanings with representational similarity analysis. *Journal of Neuroscience*, 45(19).

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. 2024. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, pages 1–11.

Jeanette A. Mumford, Tyler Davis, and Russell A. Poldrack. 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, 103:130–138.

Lee Osterhout and Phillip J Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Cathy J Price. 2012. A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. *Neuroimage*, 62(2):816–847.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2025. Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001, Abu Dhabi, UAE. Association for Computational Linguistics.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2021. Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):589–603.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561.

Mathieu Vigneau, Virginie Beaucousin, Pierre-Yves Hervé, Hugues Duffau, Fabrice Crivello, Olivier

Houde, Bernard Mazoyer, and Nathalie Tzourio-Mazoyer. 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, 30(4):1414–1432.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024. Fine-grained modeling of narrative context: A coherence perspective via retrospective questions. *arXiv preprint arXiv:2402.13551*.

Tal Yarkoni, Nicole K Speer, and Jeffrey M Zacks. 2008. Neural substrates of narrative comprehension and memory. *Neuroimage*, 41(4):1408–1425.

Shaoyun Yu, Chanyuan Gu, Kexin Huang, and Ping Li. 2024. Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science advances*, 10(21):eadn7744.

Jeffrey M Zacks, Raymond A Mar, and Navona Calarco. 2017. The cognitive neuroscience of discourse: Covered ground and new directions. In *The Routledge handbook of discourse processes*, pages 269–294. Routledge.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Yuchen Zhou, Emmy Liu, Graham Neubig, Michael J. Tarr, and Leila Wehbe. 2024. Divergences between language models and human brains. In *Advances in Neural Information Processing Systems*, volume 37, pages 137999–138031. Curran Associates, Inc.

A Details on Regions of Interest (ROIs)

The language network occupies a central role in the human brain, comprising the lateral frontal cortex (encompassing the inferior frontal gyrus and middle frontal gyrus) and the lateral temporal cortex (involving the superior and middle temporal gyri extending from the temporal pole to the posterior temporal lobe). In the majority of individuals, these regions demonstrate functional lateralization, with the left hemisphere exhibiting stronger and more spatially extensive activation during linguistic processing. Notably, emerging evidence highlights the right hemisphere's crucial contributions to language information processing, particularly manifesting distinct patterns of neural activation during higher-order linguistic functions such as pragmatic inference, metaphor comprehension, and affective prosody processing.

The network demonstrates remarkable functional specificity: The sentence > nonwords sequence contrast paradigm effectively isolates selective responses to linguistic information (particularly syntactic structuring and semantic integration) while suppressing neural engagement in non-linguistic cognitive tasks such as arithmetic reasoning, working memory, and spatial processing. This functional decoupling from the multiple-demand network (Fedorenko et al., 2024) ensures focused analysis of core discourse comprehension processes, eliminating confounding effects from perceptual encoding and executive functions. Such specificity proves particularly advantageous for analyzing texts like The Little Prince that demand deep syntactic-semantic integration due to their complex metaphorical structure and narrative architecture.

Second, the network's neural representations exhibit cross-modal and cross-linguistic universality: Large-scale neuroimaging data encompassing 45 languages (Malik-Moraleda et al., 2022) reveal stable topological organization across visual/auditory modalities and comprehension/production tasks. This domain-general consistency establishes it as an ideal framework for exploring fundamental mechanisms of human language capacity, while providing theoretical grounding for analyzing multilingual translation processing mechanisms.

Methodologically, Fedorenko's team addressed inter-individual functional-anatomical variability through probabilistic functional localization derived from fMRI data of 800 participants (Lipkin et al., 2022). Compared to traditional anatomical ROI definitions, this approach significantly enhances functional region reproducibility (inter-subject correlation coefficients >0.85), ensuring superior ecological validity. Empirical evidence confirms the network's heightened sensitivity to incremental parsing and narrative integration during natural language processing (Fedorenko et al., 2016),

effectively capturing neural signatures of advanced semantic operations like metaphor comprehension. Notably, geometric isomorphism exists between the network's neural representations and deep-layer embeddings in Transformer architectures (Tuckute et al., 2024). This representational correspondence provides crucial theoretical support for comparative analysis of processing mechanisms between biological and artificial neural networks within a unified functional framework.

B Details of General Linear Model (GLM)

The neuroimaging analysis employs a mass-univariate General Linear Model (GLM) framework to estimate task-evoked neural activity from blood-oxygen-level-dependent (BOLD) signals. The core model is expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{9}$$

where $\mathbf{Y} \in \mathbb{R}^{T\times V}$ denotes the preprocessed BOLD time series across T time points (TRs) and V voxels, $\mathbf{X} \in \mathbb{R}^{T\times P}$ the design matrix encoding experimental conditions and nuisance regressors, $\boldsymbol{\beta} \in \mathbb{R}^{P\times V}$ the unknown regression coefficients, and $\boldsymbol{\epsilon}$ the residual error with homoscedastic variance σ^2.

The design matrix $\mathbf{X}$ is constructed through the convolution of trial onsets with a canonical hemodynamic response function (HRF). For each experimental condition j with K_j events occurring at times $\{t_{\text{onset}}^{(j,k)}\}_{k=1}^{K_j}$, the corresponding regressor $\mathbf{x}_j$ is generated as:

$$\mathbf{x}_j(t) = \sum_{k=1}^{K_j} \text{HRF}(t - t_{\text{onset}}^{(j,k)}) + \sum_{m=1}^{M} \gamma_m \mathbf{n}_m(t) \tag{10}$$

where $\text{HRF}(\cdot)$ represents the double-gamma function from SPM12, and $\{\mathbf{n}_m(t)\}_{m=1}^{M}$ capture nuisance confounds (e.g., motion parameters, drift terms) weighted by coefficients γ_m.

In the present study, neural responses elicited by complex linguistic stimuli were estimated using a single-trial modeling framework based on the Least-Squares Separate (LS-S) method (Mumford et al., 2014). This analytical strategy enables sentence-level independent regression, thus facilitating a more precise dissociation of blood-oxygen-level-dependent (BOLD) signals attributable to individual trials. By modeling each target sentence separately while controlling for variance introduced by non-target trials, this approach substantially mitigates the issue of condition collinearity—a pervasive limitation in conventional block-based or condition-averaged designs. In contrast to the alternative Least-Squares All (LS-A) technique, which models all trial events concurrently, the LS-S method retains inter-trial temporal variability and improves sensitivity to transient, event-related cognitive processes through distinct parameter estimation for each trial. This increased granularity in modeling improves the reliability and interpretability of neural activation patterns associated with sentence-level processing under naturalistic language input conditions.

Parameter estimation via ordinary Least Squares-Separate(LS-S) yields voxel-wise coefficient maps:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \tag{11}$$

followed by computation of t-statistics for hypothesis testing on linear contrasts $\mathbf{c} \in \mathbb{R}^P$:

$$t_v = \frac{\mathbf{c}^\top\hat{\boldsymbol{\beta}}_v}{\sqrt{\hat{\sigma}_v^2 \cdot \mathbf{c}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{c}}}, \quad \hat{\sigma}_v^2 = \frac{\|\mathbf{Y}_v - \mathbf{X}\hat{\boldsymbol{\beta}}_v\|^2}{T - \text{rank}(\mathbf{X})} \tag{12}$$

where t_v quantifies the standardized effect size for voxel v, and $\hat{\sigma}_v^2$ estimates the residual variance. This formulation supports inference on condition-specific activation patterns while controlling for physiological noise and scanner drift.

C Details of Cross-Validated Ridge Regression for Layer-wise Encoding Analysis

The neurocognitive encoding pipeline utilizes regularized regression to quantify the semantic representations associated with specific layers of a deep neural network in brain activity patterns derived from functional magnetic resonance imaging (fMRI). This framework bridges computational neuroscience and artificial intelligence, enabling a systematic analysis of the relationship between brain responses and representations derived from deep learning models. Specifically, given the fMRI response vectors $\mathbf{y} \in \mathbb{R}^T$ from a region of interest (ROI) r and deep learning embeddings $\mathbf{X} \in \mathbb{R}^{T\times L\times D}$ (where T is the number of time points, L is the number of neural network layers, and D is the embedding dimensionality), the model's layer-wise predictive performance is computed as follows:

Region Name	Abbreviation	Functional Domain
Left Hemisphere Inferior frontal gyrus	LH_IFG	• Higher-order language processing • Lexical retrieval, syntax integration • Distributed syntactic processing
Left Hemisphere Inferior frontal gyrus orbital	LH_IFGorb	• Higher-order language processing • Lexical/sentence meaning and structure • Reading
Left Hemisphere Middle frontal gyrus	LH_MFG	• Language processing • Linguistic working memory
Left Hemisphere Anterior temporal	LH_AntTemp	• Semantic composition • Construction of complex meaning • Longer temporal integration windows
Left Hemisphere Posterior temporal	LH_PostTemp	• Core language computation "hub" • Lexical retrieval and basic semantic integration • Shorter temporal processing windows
Left Hemisphere Angular Gyrus	LH_AngG	• Semantic integration • Discourse comprehension • Interface with social cognition systems
Right Hemisphere Inferior frontal gyrus	RH_IFG	• Semantic control in sentence/narrative processing • Supports language comprehension under cognitive load
Right Hemisphere Inferior frontal gyrus orbital	RH_IFGorb	• Supports language processing under difficulty • May engage in demanding tasks
Right Hemisphere Middle frontal gyrus	RH_MFG	• Supports language processing under difficulty • May engage in demanding tasks
Right Hemisphere Anterior temporal	RH_AntTemp	• Processes social content in language • Supports language under difficulty • Overlaps with ToM-related areas
Right Hemisphere Posterior temporal	RH_PostTemp	• Processes social content in language • Integrates contextual information • Supports language under difficulty
Right Hemisphere Angular Gyrus	RH_AngG	• Supports language processing under difficulty • Sensitive to non-language ToM tasks • May be recruited for socially driven reasoning

Table 2: Brain regions, their abbreviations, and associated functional domains in language processing. *Abbreviations:* LH – Left Hemisphere; RH – Right Hemisphere; IFG – Inferior Frontal Gyrus; MFG – Middle Frontal Gyrus; AntTemp – Anterior Temporal; PostTemp – Posterior Temporal; AngG – Angular Gyrus.

$$\mathcal{L}(\mathbf{X}^{(l)}, \mathbf{y}) = \frac{1}{K} \sum_{k=1}^{K} \rho \left(\mathbf{y}_{\text{test}}^{(k)}, \hat{\mathbf{y}}_{\text{test}}^{(k)} \right) \tag{13}$$

Here, $\rho(\cdot)$ refers to the Pearson correlation coefficient, a common metric to assess the similarity between actual fMRI responses $\mathbf{y}_{\text{test}}^{(k)}$ and predicted responses $\hat{\mathbf{y}}_{\text{test}}^{(k)}$ from the ridge regression model. Ridge regression is applied to compute $\hat{\mathbf{y}}$, regularizing the regression coefficients $\boldsymbol{\beta}$ to prevent overfitting in high-dimensional datasets:

$$\hat{\boldsymbol{\beta}}^{(l)} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{X}_{\text{train}}^{(l)} \boldsymbol{\beta} - \mathbf{y}_{\text{train}} \right\|^2 + \alpha \|\boldsymbol{\beta}\|^2 \tag{14}$$

In this equation, $\mathbf{X}_{\text{train}}^{(l)}$ and $\mathbf{y}_{\text{train}}$ represent the training inputs and fMRI targets, respectively, while the regularization parameter α controls the trade-off between data fidelity and model complexity by penalizing large values of $\boldsymbol{\beta}$. This ensures the model is robust to noise and can generalize better to unseen testing data.

Feature Normalization. To ensure numerical stability and compatibility between fMRI responses $\mathbf{y}$ and model embeddings $\mathbf{X}$, all variables are z-scored prior to regression:

$$\mathbf{y} \to \frac{\mathbf{y} - \mu_y}{\sigma_y}, \quad \mathbf{X}^{(l)} \to \frac{\mathbf{X}^{(l)} - \boldsymbol{\mu}_X}{\boldsymbol{\sigma}_X} \tag{15}$$

Here, μ_y and σ_y are the mean and standard deviation of the fMRI signals across time, ensuring that $\mathbf{y}$ has zero mean and unit variance. Similarly, $\boldsymbol{\mu}_X$ and $\boldsymbol{\sigma}_X$ represent the channel-wise mean and standard deviation of the neural embeddings $\mathbf{X}^{(l)}$. This normalization step ensures that all features (both neural and brain signals) are on comparable scales, improving the stability and interpretability of the regression weights.

Layer-wise Regression Analysis. For each layer l of the neural network ($l \in \{1, ..., L\}$), we evaluate the performance of the model in predicting the fMRI responses. Ridge regression is applied independently on each layer's embeddings $\mathbf{X}^{(l)}$ to fit the training data. The prediction quality for that layer is quantified as the average Pearson correlation across K cross-validation folds:

$$\rho_l = \frac{1}{K} \sum_{k=1}^{K} \text{corr} \left(\mathbf{y}_{\text{test}}^{(k)}, \mathbf{X}_{\text{test}}^{(l,k)} \hat{\boldsymbol{\beta}}^{(l,k)} \right) \tag{16}$$

Here: - $\hat{\boldsymbol{\beta}}^{(l,k)}$ is computed from the training subset $(\mathbf{X}_{\text{train}}^{(l,k)}, \mathbf{y}_{\text{train}}^{(k)})$ using the closed-form solution:

$$\hat{\boldsymbol{\beta}}^{(l,k)} = (\mathbf{X}_{\text{train}}^{(l,k)\top} \mathbf{X}_{\text{train}}^{(l,k)} + \alpha \mathbf{I})^{-1} \mathbf{X}_{\text{train}}^{(l,k)\top} \mathbf{y}_{\text{train}} \tag{17}$$

This equation balances data fitting ($\mathbf{X}^\top \mathbf{X}$) and regularization (the identity matrix $\mathbf{I}$ scaled by α). - The hyperparameter α is selected through grid search (e.g., $\alpha \in \{10^{-3}, 10^{-2}, ..., 10^{1}\}$) optimized within a nested cross-validation loop to maximize the prediction accuracy.

This process produces a layer-wise correlation metric ρ_l, which reflects how well the embeddings at that neural layer predict ROI-specific brain responses.

Hierarchical Computational Parallelization. Given that layer-wise analysis needs to be performed across multiple **subjects**, **ROIs**, and **neural network layers**, the computational pipeline is structured as a nested series of parallel loops:

$$\underbrace{\text{Subjects}}_{\text{Outer loop}} \times \underbrace{\text{ROIs}}_{\text{Middle loop}} \times \underbrace{\text{Layers}}_{\text{Inner loop}} \to \mathcal{O}(SRL) \text{ models} \tag{18}$$

Here, S represents the number of subjects, R represents the number of ROIs analyzed, and L is the total number of neural network layers. In practice, this results in training and evaluating an extensive number of ridge regression models ($\mathcal{O}(SRL)$), thereby motivating an efficient parallel implementation.

D Input Perturbations

We show in the following the detailed input perturbations we employed for the LLMs sentence performance tasks we conducted in §4.1.

- **Option A** is the correct English translation of the Chinese sentence.
- **Option B** is constructed by scrambling the word order of Option A, thereby disrupting the syntactic structure while preserving the lexical content.
- **Option C** is formed by substituting certain words in Option A with their synonyms or words of the same part of speech, altering the surface semantics.
- **Option D** is generated by transforming the sentence structure of Option A, such as changing from active to passive voice, while maintaining the original meaning.

- **Option E** is created by inserting or deleting information in Option A, thus introducing semantic noise or omitting key details.

E Details on the LLMs

We present the detailed information of the 14 LLMs in Table 3. During the experiments, we used the `transformers` (Wolf et al., 2020) and `pytorch` (Paszke et al., 2019) library for training the models. All experiments were conducted on an NVIDIA® A100 80 GB RAM GPU.

Scaling	Model	Layers
~110M	Bert-base-uncased	12
~7B	Qwen2.5-7B	28
	Qwen2.5-7B-Instruct	28
	Mistral-7B-v0.3	32
	Mistral-7B-Instruct-v0.3	32
	Baichuan2-7B-Chat	32
	DeepSeek-R1-Distill-Qwen-7B	32
	opt-6.7b	32
~8B	Llama-3.1-8B	32
	Llama-3.1-8B-Instruct	32
~9B	gemma-2-9b	28
	gemma-2-9b-it	28
	Glm-4-9b-chat-hf	32
	Glm-4-9b-hf	32

Table 3: Layer numbers of transformer-based models across different parameter scaling levels. Model names link to their Hugging Face repositories.

F Additional Experimental Results

Model Performance and Activation Correlation. To comprehensively assess the correspondence between computational models and brain activity, we compared model performance across multiple brain regions and layers, as addition to §4.2. Figure 9 displays the distribution of correlation coefficients for 12 bilateral regions of interest (ROIs), allowing for a direct comparison of model prediction accuracy across distinct brain areas.

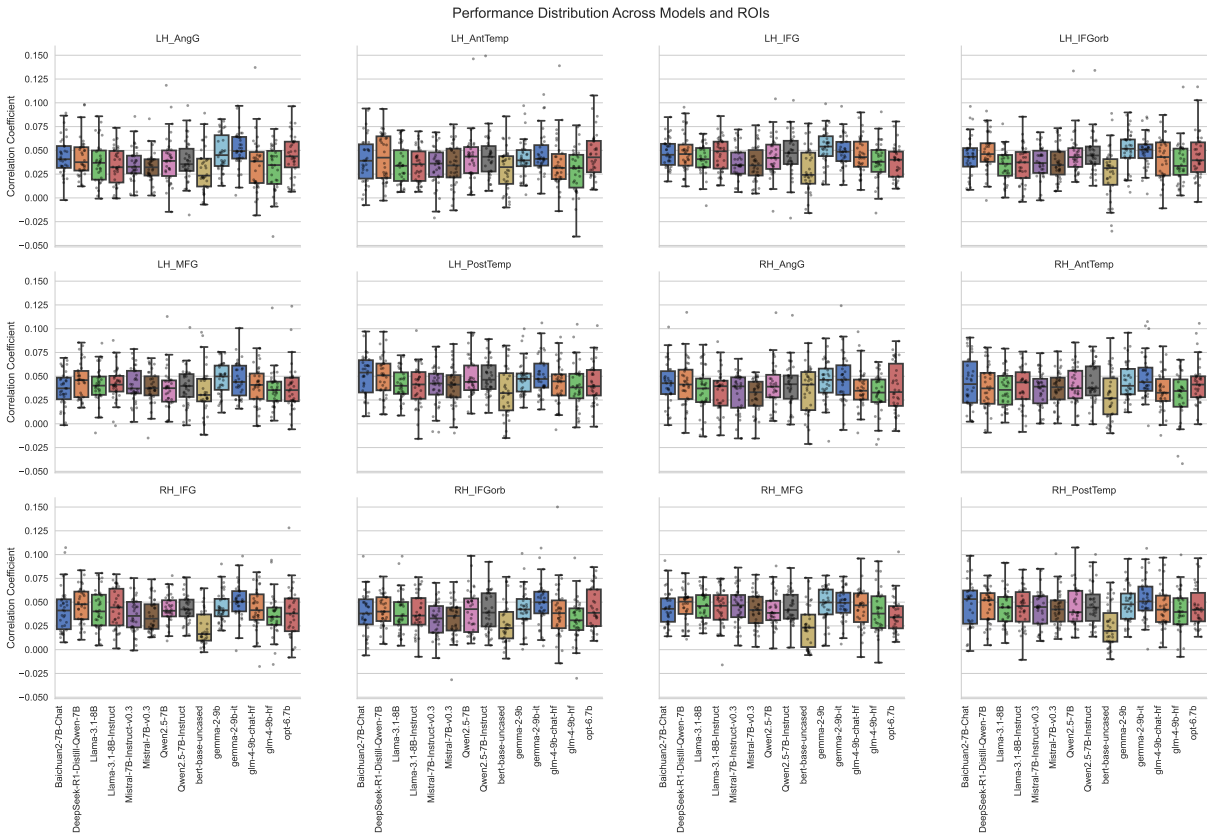

Figure 9: Multi-model performance distribution across bilateral brain regions, showing correlation coefficients (range: -0.05 to 0.15) for 12 ROIs in 4×3 grid layout